

Further Theoretical Study of Distribution Separation Method for Information Retrieval

Peng Zhang¹, Qian Yu^{1,2}, Yuexian Hou^{1,*}, Dawei Song^{1,3,*}, Jingfei Li¹, and Bin Hu⁴

¹Tianjin University, China

²The Hong Kong Polytechnic University, Hong Kong

³The Open University, United Kingdom

⁴Lanzhou University, China

*{yxhou,dwsong}@tju.edu.cn

ABSTRACT

Recently, a Distribution Separation Method (DSM) is proposed for relevant feedback in information retrieval, which aims to approximate the true relevance distribution by separating a seed irrelevance distribution from the mixture one. While DSM achieved a promising empirical performance, theoretical analysis of DSM is still need further study and comparison with other relative retrieval model. In this article, we first generalize DSM's theoretical property, by proving that its minimum correlation assumption is equivalent to the maximum (original and symmetrized) KL-Divergence assumption. Second, we also analytically show that the EM algorithm in a well-known Mixture Model is essentially a distribution separation process and can be simplified using the linear separation algorithm in DSM. Some empirical results are also presented to support our theoretical analysis.

Introduction

Relevant feedback is an effective method in information retrieval, which can significantly improve the retrieval performance. However, the approximation of relevance model is usually still a mixture model containing an irrelevant component. A Distribution Separation Method is recently proposed to solve this problem.¹ The formulation of the basic DSM was based on two assumptions, namely the linear combination assumption and minimum correlation assumption. The former assumes that the mixture term distribution is a linear combination of the relevance and irrelevance distributions, while the later assumes that the relevance distribution should have the minimum correlation with the irrelevance distribution. The basic DSM provided a lower bound analysis for the linear combination coefficient, based on which the desired relevance distribution can be estimated. It was also proved that the lower bound of the linear combination coefficient corresponds to the condition of the minimum Pearson correlation coefficient between DSM's output relevance distribution and an input seed irrelevance distribution.

In this article, we theoretically extend the generality of the aforementioned linear combination analysis and the minimum correlation analysis of DSM. First, we propose to explore the effect of DSM on the KL-divergence between DSM's output distribution and the seed irrelevance distribution. We theoretically prove that the lower-bound analysis can also be applied to KL-divergence, and the minimum correlation coefficient corresponds to the maximum KL-divergence. We further prove that the decreasing correlation coefficient leads to a maximum symmetrized KL-divergence as well as JS-divergence.

Second, we investigate the relations between DSM and a well-known Mixture Model Feedback (MMF) approach² in information retrieval. We theoretically show that the EM-based iterative algorithm in MMF is essentially a distribution separation process and thus its iterative steps can be simplified by the linear separation technique developed in DSM without decline of performance.

Basic Analysis of DSM

In this section, we briefly describes assumptions and analysis of the basic DSM.¹ We use M to represent the mixture term distribution derived from all the feedback documents, and we believe that M is a mixture of relevance term distribution R and irrelevance term distribution I . In addition, we assume that only part of the irrelevance distribution I_S (also called as seed irrelevance distribution) is available, while the other part of irrelevance distribution is unknown (denoted as $I_{\bar{S}}$).

The task of DSM can be defined as follows: given the mixture distribution M and a seed irrelevance distribution I_S , to derive an output distribution that can approximates the R as closely as possible. Specifically, as shown in Figure 1, the task of DSM can be divided into two problems: (1) How to separate I_S from M , and derive $l(R, I_{\bar{S}})$, which is less noisy but is still a mixture of the true relevance distribution (R) and the unknown irrelevance distribution ($I_{\bar{S}}$). (2) How to further refine the

Notation	Description
M	Mixture term distribution
R	Relevance term distribution
I	Irrelevance term distribution.
I_S	Seed Irrelevance distribution
$I_{\bar{S}}$	Unknown Irrelevance distribution
$F(i)$	Probability of the i^{th} term in any distribution F
$l(F, G)$	Linear combination of distributions F and G

Table 1. Notations

$$M = l(l(R, I_{\bar{S}}), I_S)$$

+		-			-	
+	+	-	-	-	-	-
R		$I_{\bar{S}}$			I_S	
+	+	-	-	-	-	-

$$l(R, I_{\bar{S}})$$

Figure 1. An illustration of the linear combination $l(\cdot, \cdot)$ between two term distributions.

derived distribution $l(R, I_{\bar{S}})$ to approximate R as closely as possible? To solve the above two problems, DSM assumes that a term distribution derived from a feedback document set D is a linear combination of two term distributions, which are derived from two document subsets that form a partition of D . Under such a condition, the mixture distribution M derived from all the feedback documents can be a linear combination of R (derived from relevant documents) and I (derived from irrelevant documents). As shown in Figure 1, M can be a linear combination of two distributions I_S and $l(R, I_{\bar{S}})$, where $l(R, I_{\bar{S}})$ is also a linear combination of R and $I_{\bar{S}}$. Bear in mind that in the linear combination of R and I , both R and I are unknown for DSM. Therefore, determining the linear combination of I_S and $l(R, I_{\bar{S}})$ and the linear combination of R and $I_{\bar{S}}$ are key for solving the above problem (1) and (2), respectively.

Linear Combination Analysis

Since M is a nested linear combination $l(l(R, I_{\bar{S}}), I_S)$, it can be represented as:

$$M = \lambda \times l(R, I_{\bar{S}}) + (1 - \lambda) \times I_S \quad (1)$$

where λ ($0 < \lambda \leq 1$) is the real linear coefficient. The problem of estimating $l(R, I_{\bar{S}})$ does not have a unique solution generally with coefficient λ is unknown. Therefore, the key is to estimate λ . Let $\hat{\lambda}$ ($0 < \hat{\lambda} \leq 1$) denote an estimate of λ , and correspondingly let $\hat{l}(R, I_{\bar{S}})$ be the estimation of the desired distribution $l(R, I_{\bar{S}})$. According to Eq. 1, we have

$$\hat{l}(R, I_{\bar{S}}) = \frac{1}{\hat{\lambda}} \times M + (1 - \frac{1}{\hat{\lambda}}) \times I_S. \quad (2)$$

Then with the constraint that all the values in the distribution $\hat{l}(R, I_{\bar{S}})$ are not less than 0, we can have

$$\hat{\lambda} \times \mathbf{1} \succcurlyeq (\mathbf{1} - M./I_S) \quad (3)$$

where $\mathbf{1}$ stands for a vector in which all the entries are 1, and $./$ denotes the entry-by-entry division of M by I_S . Note that if there is zero value in I_S , then $\hat{\lambda} > 1 - \infty$. It is still valid since $\hat{\lambda} > 0$. Effectively, Eq. 3 sets a lower bound of $\hat{\lambda}$:

$$\lambda_L = \max(\mathbf{1} - M./I_S) \quad (4)$$

where $\max(\cdot)$ denotes the max value in the resultant vector $\mathbf{1} - M./I_S$. The lower bound λ_L itself also determines an estimate of $l(R, I_{\bar{S}})$, denoted as $l_L(R, I_{\bar{S}})$.

The lower bound λ_L is essential to the estimation of λ . Now, we present an important property of λ_L in Lemma 1. For simplicity, we use some simplified notations listed in Table 2. Lemma 1 guarantees that if there exists zero value (e.g., for a

Original	Simplified	Linear Coefficient
$l(R, I_{\bar{S}})(i)$	$l(i)$	λ
$\hat{l}(R, I_{\bar{S}})(i)$	$\hat{l}(i)$	$\hat{\lambda}$ (estimate of λ)
$l_L(R, I_{\bar{S}})(i)$	$l_L(i)$	λ_L (lower bound of $\hat{\lambda}$)

Table 2. Simplified Notations

term i , $l(i) = 0$) in $l(R, I_{\bar{S}})$, then $\lambda = \lambda_L$. In relevance feedback, if there is no distribution smoothing step involved for feedback model, zero values often exist in $l(R, I_S)$, leading to the distribution $l_L(R, I_{\bar{S}})$ w.r.t. λ_L being exactly the desired distribution $l(R, I_{\bar{S}})$ w.r.t. λ .

Lemma 1. *If there exists a zero value in $l(R, I_{\bar{S}})$, then $\lambda = \lambda_L$, leading to $l(R, I_{\bar{S}}) = l_L(R, I_{\bar{S}})$.*

The proof can be found in.¹ In the IR background, after applying smoothing (usually with the collection model), there would be no zero values, but instead a lot of small values exist, in $l(R, I_{\bar{S}})$. In this case, Remark 1 guarantees the approximate equality between λ_L and λ . The detailed description of this remark can be found in.¹

Remark 1. *If there is no zero value, but there exist a few very small values in $l(R, I_{\bar{S}})$, i.e., $0 < l(i) \leq \delta$, where δ is a very small value, then $l_L(R, I_{\bar{S}})$ will be approximately equal to $l(R, I_{\bar{S}})$.*

Minimum Correlation Analysis

In this section, we go in-depth to see another property of the combination coefficient and its lower bound. Specifically, we analyze the correlation between $\hat{l}(R, I_{\bar{S}})$ and I_S , along with the decreasing coefficient $\hat{\lambda}$. Pearson product-moment correlation coefficient,³ denoted as ρ ($-1 \leq \rho \leq 1$), is used as the correlation measurement.

Proposition 1. *If $\hat{\lambda}$ ($\hat{\lambda} > 0$) decreases, the correlation coefficient between $\hat{l}(R, I_{\bar{S}})$ and I_S , i.e., $\rho(\hat{l}(R, I_{\bar{S}}), I_S)$, will decrease.*

More precisely, the condition of $\hat{\lambda}$ is $\lambda_L \leq \hat{\lambda} \leq 1$ to ensure that the output distribution $\hat{l}(R, I_{\bar{S}})$ is a distribution. The proof of Proposition 1 can be found in.¹ According to Proposition 1, among all $\hat{\lambda} \in [\lambda_L, 1]$, λ_L corresponds to $\min(\rho)$, i.e., the minimum correlation coefficient between $\hat{l}(R, I_{\bar{S}})$ and I_S .

We can also change the *minimum correlation coefficient* (i.e., $\min(\rho)$) to *minimum squared correlation coefficient* (i.e., $\min(\rho^2)$). This idea can be formulated as the following optimization problem:

$$\min_{\hat{\lambda}} [\rho(\hat{l}(R, I_{\bar{S}}), I_S)]^2 \quad \text{s.t.} \quad \lambda_L \leq \hat{\lambda} \leq 1 \quad (5)$$

To solve this optimization problem, we need to first find a $\hat{\lambda}$ such that the corresponding $\rho(\hat{l}(R, I_{\bar{S}}), I_S) = 0$. According to the proof of Proposition 1 in,¹ this $\hat{\lambda} = -\frac{a}{b}$, where $a = \sum_i^m (I_S(i) - \frac{1}{m})(M(i) - I_S(i))$, $b = \sum_i^m (I_S(i) - \frac{1}{m})^2$, and m is the number of terms. Then, we need to check whether $\lambda_L \leq -\frac{a}{b} \leq 1$ holds. If it holds, the optimal linear coefficient $\hat{\lambda}$ for the optimization problem in Eq. 5 is $-\frac{a}{b}$. Otherwise, we just compare the values of $[\rho(\hat{l}(R, I_{\bar{S}}), I_S)]^2$ w.r.t. $\hat{\lambda} = 1$ and $\hat{\lambda} = \lambda_L$, in order to get the optimal $\hat{\lambda}$ of the objective function in Eq. 5.

Generalized Analysis of DSM

Now, we generalize DSM's minimum correlation assumption, by extending the minimum correlation analysis to the analysis of the maximum KL-divergence, the maximum symmetrized KL-divergence and the maximum JS-divergence.

Effect of DSM on Maximizing KL-Divergence

Recall that in Section , Proposition 1 shows that after the distribution separation process, the Pearson correlation coefficient between DSM's output distribution $\hat{l}(R, I_{\bar{S}})$ and the seed irrelevance distribution I_S can be minimized. Here, we further analyze the effect of DSM on the KL-divergence between $\hat{l}(R, I_{\bar{S}})$ and I_S .

Specifically, we propose the following Proposition 2, which proves that if $\hat{\lambda}$ decreases, the KL-divergence between $\hat{l}(R, I_{\bar{S}})$ and I_S will be increased monotonously.

Proposition 2. *If $\hat{\lambda}$ ($\hat{\lambda} > 0$) decreases, the KL-divergence between $\hat{l}(R, I_{\bar{S}})$ and I_S will increase.*

Proof. Using the simplified notations in Table 2, let the KL-divergence of between $\hat{l}(R, I_{\bar{S}})$ and I_S be formulated as

$$D(\hat{l}(R, I_{\bar{S}}), I_S) = \sum_{i=1}^m \hat{l}(R, I_{\bar{S}})(i) \log\left(\frac{\hat{l}(R, I_{\bar{S}})(i)}{I_S(i)}\right) = \sum_{i=1}^m \hat{l}(i) \log\left(\frac{\hat{l}(i)}{I_S(i)}\right) \quad (6)$$

Now let $\xi = 1/\hat{\lambda}$ as we did in the proof of Proposition 1 (see¹). According to Eq. 2, we have $\hat{l}(R, I_{\bar{S}}) = \xi \times M + (1 - \xi) \times I_S$. It then turns out that

$$\hat{l}(i) = \xi \times (M(i) - I_S(i)) + I_S(i). \quad (7)$$

Based on Eq. 6 and 7, we get

$$D(\hat{l}(R, I_{\bar{S}}), I_S) = \sum_{i=1}^m (\xi \times (M(i) - I_S(i)) + I_S(i)) \log\left(\frac{\xi \times (M(i) - I_S(i)) + I_S(i)}{I_S(i)}\right) \quad (8)$$

Let $D(\xi) = D(\hat{l}(R, I_{\bar{S}}), I_S)$. The derivative of $D(\xi)$ can be calculated as

$$D'(\xi) = \sum_{i=1}^m [M(i) - I_S(i) + (M(i) - I_S(i)) \log\left(\frac{\xi \times (M(i) - I_S(i)) + I_S(i)}{I_S(i)}\right)] \quad (9)$$

Since $\sum_{i=1}^m M(i) = 1$ and $\sum_{i=1}^m I_S(i) = 1$, $\sum_{i=1}^m [M(i) - I_S(i)]$ becomes 0. We then have

$$\begin{aligned} D'(\xi) &= \sum_{i=1}^m (M(i) - I_S(i)) \log\left(\frac{\xi \times (M(i) - I_S(i)) + I_S(i)}{I_S(i)}\right) \\ &= \sum_{i=1}^m (M(i) - I_S(i)) \log\left(\frac{\xi \times (M(i) - I_S(i))}{I_S(i)} + 1\right) \end{aligned} \quad (10)$$

Let the i^{th} term in the summation of Eq. 10 is

$$D'(\xi)(i) = (M(i) - I_S(i)) \log\left(\frac{\xi \times (M(i) - I_S(i))}{I_S(i)} + 1\right)$$

It turns out that when $M(i) > I_S(i)$ and $M(i) < I_S(i)$, $D'(\xi)(i)$ is greater than 0. When $M(i) = I_S(i)$, $D'(\xi)(i)$ is 0. However, $M(i)$ not always equals to $I_S(i)$. Therefore, $D'(\xi) = \sum_{i=1}^m D'(\xi)(i)$ is greater than 0.

In conclusion, we have $D'(\xi) > 0$. It means that $D(\xi)$ (i.e., $D(\hat{l}(R, I_{\bar{S}}), I_S)$) increases after ξ increases. Since $\lambda = 1/\xi$, after $\hat{\lambda}$ decreases, $D(\hat{l}(R, I_{\bar{S}}), I_S)$ will increase. \square

According to Proposition 2, if λ reduced to its lower bound λ_L , then the corresponding KL-divergence $D(I_L(R, I_{\bar{S}}), I_S)$ will be the maximum value for all the legal $\hat{\lambda}$ ($\lambda_L \leq \hat{\lambda} < 1$). In this case, the output distribution of DSM will have the maximum KL-divergence with the seed irrelevance distribution.

Effect of DSM on Maximizing symmetrized KL-Divergence

Having shown the effect of reducing the coefficient $\hat{\lambda}$ on the KL-divergence between $\hat{l}(R, I_{\bar{S}})$ and I_S , we now investigate the effect on the symmetrized KL-divergence between two involved distributions by proving the following proposition.

Proposition 3. *If $\hat{\lambda}$ ($\hat{\lambda} > 0$) decreases, the symmetrized KL-divergence between $\hat{l}(R, I_{\bar{S}})$ and I_S will increase.*

Proof. Let the symmetrized KL-divergence of between $\hat{l}(R, I_{\bar{S}})$ and I_S be denoted as

$$SD(\hat{l}(R, I_{\bar{S}}), I_S) = D(\hat{l}(R, I_{\bar{S}}), I_S) + D(I_S, \hat{l}(R, I_{\bar{S}})) \quad (11)$$

Since we have proved in Proposition 2 that the increasing trend of $D(\hat{l}(R, I_{\bar{S}}), I_S)$ when $\hat{\lambda}$ decreases, we now only need to prove the same result for $D(I_S, \hat{l}(R, I_{\bar{S}}))$, which is computed by:

$$D(I_S, \hat{l}(R, I_{\bar{S}})) = \sum_i I_S(i) \log\left(\frac{I_S(i)}{\hat{l}(i)}\right) = \sum_i I_S(i) \log I_S(i) - \sum_i I_S(i) \log \hat{l}(i) \quad (12)$$

Now let $\xi = 1/\hat{\lambda}$. According to Eq. 2, we have $\hat{l}(R, I_S) = \xi \times M + (1 - \xi) \times I_S$. It then turns out that

$$\hat{l}(i) = \xi \times (M(i) - I_S(i)) + I_S(i). \quad (13)$$

Based on Eq. 12 and Eq. 13, we get:

$$D(I_S, \hat{l}(R, I_S)) = \sum_i I_S(i) \log I_S(i) - \sum_i I_S(i) \log(\xi \times (M(i) - I_S(i)) + I_S(i)) \quad (14)$$

Let $D(\xi) = D(I_S, \hat{l}(R, I_S))$. The derivative of $D(\xi)$ can be calculated as

$$D'(\xi) = \sum_i \frac{-I_S(i)(M(i) - I_S(i))}{\xi \times (M(i) - I_S(i)) + I_S(i)} = \sum_i \frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} \quad (15)$$

Since $M(i)$ is a linear combination of $\hat{l}(i)$ and $I_S(i)$, $M(i)$ is a in-between value of $\hat{l}(i)$ and $I_S(i)$. In other words, if $M(i) > I_S(i)$, then $\hat{l}(i) > M(i) > I_S(i)$, while $\hat{l}(i) < M(i) < I_S(i)$ if $M(i) < I_S(i)$.

If $M(i) > I_S(i)$, since $M(i) - I_S(i) > 0$ and $0 < \frac{I_S(i)}{\hat{l}(i)} < 1$, we have

$$\frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} > -(M(i) - I_S(i))$$

If $M(i) < I_S(i)$, since $M(i) - I_S(i) < 0$ and $\frac{I_S(i)}{\hat{l}(i)} > 1$, we have

$$\frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} > -(M(i) - I_S(i))$$

We then have

$$\begin{aligned} D'(\xi) &= \sum_i \frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} \\ &= \sum_{i:M(i) > I_S(i)} \frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} + \sum_{i:M(i) < I_S(i)} \frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} + \sum_{i:M(i) = I_S(i)} \frac{-I_S(i)(M(i) - I_S(i))}{\hat{l}(i)} \\ &> \sum_{i:M(i) > I_S(i)} -(M(i) - I_S(i)) + \sum_{i:M(i) < I_S(i)} -(M(i) - I_S(i)) + \sum_{i:M(i) = I_S(i)} -(M(i) - I_S(i)) \\ &= \sum_i -(M(i) - I_S(i)) \\ &= 0 \end{aligned} \quad (16)$$

We now have $D'(\xi) > 0$. It means that $D(\xi)$ (i.e., $D(I_S, \hat{l}(R, I_S))$) will increase after ξ increases. Since $\lambda = 1/\xi$, after $\hat{\lambda}$ decreases, $D(I_S, \hat{l}(R, I_S))$ will increase. Combined with the result proved in Proposition 2, we can conclude that when $\hat{\lambda}$ decreases, the symmetrized KL-divergence $D(\hat{l}(R, I_S), I_S) + D(I_S, \hat{l}(R, I_S))$ will increase monotonically. \square

According to Proposition 3, if λ reduced to its lower bound λ_L , then the corresponding symmetrized KL-divergence $D(I_S, \hat{l}(R, I_S))$ will be the maximum value. In this case, the output distribution of DSM will have the maximum symmetrized KL-divergence with the seed irrelevance distribution.

Effect of DSM on Maximizing JS-Divergence

Now, let us further study the reduction of the coefficient $\hat{\lambda}$ on its role in maximizing the JS-divergence between DSM's output distribution $\hat{l}(R, I_S)$ and the seed irrelevance distribution I_S .

Proposition 4. *If $\hat{\lambda}$ ($\hat{\lambda} > 0$) decreases, the JS-divergence between $\hat{l}(R, I_S)$ and I_S will increase.*

Proof. Let the JS-divergence of between $\hat{l}(R, I_S)$ and I_S be denoted as

$$JS(\hat{l}(R, I_S), I_S) = \frac{1}{2} (D(\hat{l}(R, I_S), \frac{\hat{l}(R, I_S) + I_S}{2}) + D(I_S, \frac{\hat{l}(R, I_S) + I_S}{2})) \quad (17)$$

Now let $\xi = 1/\hat{\lambda}$. Based on Eq. 17 and Eq. 13, we get

$$JS(\hat{l}(R, I_S), I_S) = \frac{1}{2} \sum_i (\xi \times (M(i) - I_S(i)) + I_S(i)) \log \left(\frac{2\xi \times (M(i) - I_S(i)) + 2I_S(i)}{\xi \times (M(i) - I_S(i)) + 2I_S(i)} \right) + \frac{1}{2} \sum_i I_S(i) \log \left(\frac{2I_S(i)}{\xi \times (M(i) - I_S(i)) + 2I_S(i)} \right) \quad (18)$$

Let $J(\xi) = 2 \times JS(\hat{l}, I_S)$, we can have

$$J(\xi) = \sum_i (\xi \times (M(i) - I_S(i)) + I_S(i)) \log(2\xi \times (M(i) - I_S(i)) + 2I_S(i)) - \sum_i (\xi \times (M(i) - I_S(i)) + I_S(i)) \log(\xi \times (M(i) - I_S(i)) + 2I_S(i)) + \sum_i I_S(i) \log 2I_S(i) - \sum_i I_S(i) \log(\xi \times (M(i) - I_S(i)) + 2I_S(i)) \quad (19)$$

The derivative of $J(\xi)$ can be calculated as

$$J'(\xi) = \sum_i (M(i) - I_S(i)) \log(2\xi \times (M(i) - I_S(i)) + 2I_S(i)) + \sum_i (\xi \times (M(i) - I_S(i)) + I_S(i)) \frac{M(i) - I_S(i)}{\xi \times (M(i) - I_S(i)) + I_S(i)} - \sum_i (M(i) - I_S(i)) \log(\xi \times (M(i) - I_S(i)) + 2I_S(i)) - \sum_i (\xi \times (M(i) - I_S(i)) + I_S(i)) \frac{M(i) - I_S(i)}{\xi \times (M(i) - I_S(i)) + 2I_S(i)} - \sum_i I_S(i) \frac{M(i) - I_S(i)}{\xi \times (M(i) - I_S(i)) + 2I_S(i)} \quad (20)$$

Since $\hat{l}(i) = \xi \times (M(i) - I_S(i)) + I_S(i)$ (see Eq. 13), we have

$$J'(\xi) = \sum_i (M(i) - I_S(i)) \log(2\hat{l}(i)) + \sum_i \hat{l}(i) \frac{M(i) - I_S(i)}{\hat{l}(i)} - \sum_i (M(i) - I_S(i)) \log(\hat{l}(i) + I_S(i)) - \sum_i \hat{l}(i) \frac{M(i) - I_S(i)}{\hat{l}(i) + I_S(i)} - \sum_i I_S(i) \frac{M(i) - I_S(i)}{\hat{l}(i) + I_S(i)} = \sum_i (M(i) - I_S(i)) \log \left(\frac{2\hat{l}(i)}{\hat{l}(i) + I_S(i)} \right) - \sum_i (\hat{l}(i) + I_S(i)) \frac{M(i) - I_S(i)}{\hat{l}(i) + I_S(i)} = \sum_i (M(i) - I_S(i)) \log \left(\frac{2\hat{l}(i)}{\hat{l}(i) + I_S(i)} \right) \quad (21)$$

If $M(i) > I_S(i)$, since $M(i) - I_S(i) > 0$ and $\hat{l}(i) > I_S(i) \geq 0$, we have

$$(M(i) - I_S(i)) \log \left(\frac{2\hat{l}(i)}{\hat{l}(i) + I_S(i)} \right) > 0$$

If $M(i) < I_S(i)$, since $M(i) - I_S(i) < 0$ and $0 \leq \hat{l}(i) < I_S(i)$, we have

$$(M(i) - I_S(i)) \log \left(\frac{2\hat{l}(i)}{\hat{l}(i) + I_S(i)} \right) > 0$$

We then have $J'(\xi) > 0$. It means that $JS(\hat{l}(R, I_S), I_S)$ increases after ξ increases. Since $\lambda = 1/\xi$, after $\hat{\lambda}$ decreases, $JS(\hat{l}(R, I_S), I_S)$ will increase monotonically. \square

According to Proposition 4, if λ reduced to its lower bound λ_L , then the corresponding JS-divergence $JS(\hat{l}(R, I_S), I_S)$ will be the maximum value. In this case, the output distribution of DSM will have the maximum symmetrized JS-divergence with the seed irrelevance distribution.

Analysis of Relation between DSM and Mixture Model Feedback

A related model of DSM is the Mixture Model Feedback (MMF) approach which assumes that feedback documents are generated from a mixture model with two multinomial components, i.e., the query topic model and the collection model.² In this section we theoretically investigate the linear combination condition of DSM in a related Mixture Model.²

The estimation of the output “relevant” query model of MMF is trying to purify the feedback document by eliminating the effect of the collection model, since the collection model contains background noise which can be regarded as the “irrelevant” content in the feedback document.² In this sense, similar to DSM, the task of MMF can also be regarded as removing the irrelevant part in the mixture model. However, to our knowledge, researchers have not investigated if the linear combination assumption is valid or not in MMF. We will theoretically prove that the mixture model in MMF is indeed a linear combination of “relevant” and “irrelevant” parts. This theoretical result can lead to a simplified version of MMF based on linear separation algorithm of DSM. Next, we first review the Mixture Model Feedback Approach in detail.

Review of Mixture Model Feedback Approach

In the Mixture Model Feedback (MMF) approach, the likelihood of feedback documents (\mathcal{F}) can be written as:

$$\log p(\mathcal{F}|\theta_F) = \sum_{d \in \mathcal{F}} \sum_{w \in d} c(w; d) \log[\lambda p(w|\theta_F) + (1 - \lambda)p(w|C)] \quad (22)$$

where $c(w; d)$ is the count of a term w in a document d , $p(w|\theta_F)$ is the query topic model which can be regarded as the relevance distribution to be estimated, and $p(w|C)$ is the collection distribution containing the background information. The empirically assigned parameter λ is the amount of true relevance distribution and $1 - \lambda$ indicates the amount of background noise, i.e., the influence of C in the feedback documents. An EM method² is developed to estimate the relevance distribution via maximizing the likelihood in Equation 22. It contains iterations of two steps:⁴

$$p(z_w = 1|\mathcal{F}, \theta_F^{(n)}) = \frac{(1 - \lambda)p(w|C)}{\lambda p(w|\theta_F^{(n)}) + (1 - \lambda)p(w|C)} \quad E \text{ step} \quad (23)$$

$$p(w|R^{(n+1)}) = \frac{\sum_{d \in \mathcal{F}} (1 - p(z_w = 1|\mathcal{F}, \theta_F^{(n)}))c(w, d)}{\sum_{d \in \mathcal{F}} \sum_{w^* \in V} (1 - p(z_{w^*} = 1|\mathcal{F}, \theta_F^{(n)}))c(w^*, d)} \quad M \text{ step} \quad (24)$$

where $p(z_w = 1|\mathcal{F}, \theta_F^{(n)})$ is the probability that the word w is from background distribution, given the current estimation of relevance distribution ($\theta_F^{(n)}$). This estimation can be regarded as a procedure to obtain relevant information from feedback documents while filtering the influence of collection distribution, leading to a more discriminative relevance model. It should be noted that in Eq. 22, due to the log operator within the summations (i.e., $\sum_{d \in \mathcal{F}} \sum_{w \in d} c(w; d)$), it does not directly show that the mixture model is a linear combination of the collection model and the query topic model. Therefore, an EM algorithm is adopted to estimate the query topic model θ_F .

The Simplification of EM Algorithm in MMF via Linear Separation Algorithm

Now, we first explore the connections between DSM and MMF. Once λ is given (either by the estimation in DSM or by an assigned value in MMF), the next step is to estimate the true relevance distribution R . We will demonstrate that if the EM algorithm (in MMF) converges, the mixture model of the feedback documents is a linear combination of the collection model and the output model of the EM iterative algorithm.

Proposition 5. *If the EM algorithm (in MMF) converges, the mixture model of the feedback documents is a linear combination of the collection model and the output relevance model of the EM iterative algorithm.*

Proof. When the EM method converges in the mixture model feedback approach, without loss of generality, we let $p(w|\theta_F^{(n+1)}) = p(w|\theta_F^{(n)})$. In addition to this, we can replace the $p(z_w = 1|\mathcal{F}, \theta_F^{(n)})$ in Equation 24 using Equation 23:

$$\begin{aligned} p(w|\theta_F^{n+1}) &= p(w|\theta_F^{(n)}) = \frac{\sum_{d \in \mathcal{F}} \left[1 - \frac{(1-\lambda)p(w|C)}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} \right] \cdot c(w, d)}{\sum_{d \in \mathcal{F}} \sum_{w^* \in V} \left[1 - \frac{(1-\lambda)p(w^*|C)}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \right] \cdot c(w^*, d)} \\ &= \frac{\sum_{d \in \mathcal{F}} \frac{\lambda p(w|\theta_F^{(n)})}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} \cdot c(w, d)}{\sum_{d \in \mathcal{F}} \sum_{w^* \in V} \frac{\lambda p(w^*|\theta_F^{(n)})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \cdot c(w^*, d)} \end{aligned} \quad (25)$$

By dividing $p(w|\theta_F^{(n)})$ in both the second term and the fourth term in Eq. 25, we get:

$$1 = \frac{\sum_{d \in \mathcal{F}} \frac{\lambda}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} \cdot c(w, d)}{\sum_{d \in \mathcal{F}} \sum_{w^* \in V} \frac{\lambda p(w^*|\theta_F^{(n)})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \cdot c(w^*, d)} \quad (26)$$

Then, it turns out that, for a particular word w :

$$\sum_{d \in \mathcal{F}} \frac{c(w, d)}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} = \sum_{d \in \mathcal{F}} \sum_{w^* \in V} \frac{p(w^*|\theta_F^{(n)})c(w^*, d)}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)}$$

Replace $\sum_{d \in \mathcal{F}} c(w, d_i)$ with $c(w^*, \mathcal{F})$, we can get:

$$\frac{c(w, \mathcal{F})}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} = \sum_{w^* \in V} \frac{p(w^*|\theta_F^{(n)})c(w^*, \mathcal{F})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \quad (27)$$

If each side of Equation 27 is multiplied by $(1-\lambda)p(w|C)$ then it becomes:

$$\frac{(1-\lambda)p(w|C)c(w, \mathcal{F})}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} = (1-\lambda)p(w|C) \sum_{w^* \in V} \frac{p(w^*|\theta_F^{(n)})c(w^*, \mathcal{F})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \quad (28)$$

We can obtain the Equation 28 for any word w in the vocabulary, and now we sum them together:

$$\sum_{w^* \in V} \frac{(1-\lambda)p(w^*|C)c(w^*, \mathcal{F})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} = \sum_{w^* \in V} \frac{(1-\lambda)p(w^*|\theta_F^{(n)})c(w^*, \mathcal{F})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \quad (29)$$

then we add $\sum_{w^* \in V} \frac{\lambda p(w^*|\theta_F^{(n)})c(w^*, \mathcal{F})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)}$ to both sides of Equation 29:

$$\sum_{w^* \in V} c(w^*, \mathcal{F}) = \sum_{w^* \in V} \frac{p(w^*|\theta_F^{(n)})c(w^*, \mathcal{F})}{\lambda p(w^*|\theta_F^{(n)}) + (1-\lambda)p(w^*|C)} \quad (30)$$

According to Equation 27 and Equation 30, we get:

$$\frac{c(w, \mathcal{F})}{\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C)} = \sum_{w^* \in V} c(w^*, \mathcal{F}) \quad (31)$$

namely,

$$\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C) = \frac{c(w, \mathcal{F})}{\sum_{w^* \in V} c(w^*, \mathcal{F})} = tf(w, \mathcal{F}) \quad (32)$$

where $tf(w, \mathcal{F})$ is the term frequency in the feedback documents. The above equation illustrates that the estimated distribution $\theta_F^{(n)}$ fits in a linear combination as used in Eq. 1 of DSM. \square

This proposition actually can be formulated as:

$$\lambda p(w|\theta_F^{(n)}) + (1-\lambda)p(w|C) = \frac{c(w, \mathcal{F})}{\sum_{w^* \in V} c(w^*, \mathcal{F})} = tf(w, \mathcal{F}) \quad (33)$$

where $tf(w, \mathcal{F})$ is the mixture model which represents the term frequency in the feedback documents, $p(w|C)$ is the collection model and $p(w|\theta_F^{(n)})$ is the estimated relevance model output by the n^{th} step of the EM iterative algorithm in MMF. The above equation can be derived as:

$$p(w|\theta_F^{(n)}) = \frac{1}{\lambda} \cdot tf(w, \mathcal{F}) + \left(1 - \frac{1}{\lambda}\right) \cdot p(w|C) \quad (34)$$

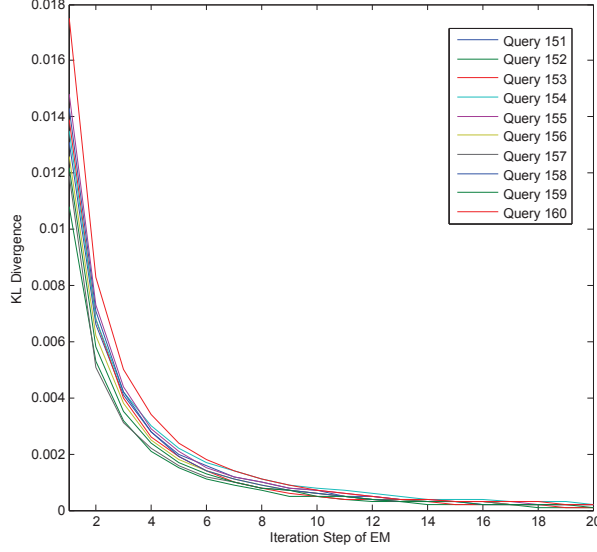


Figure 2. KL Divergence between result of Equation 2 and the current estimation of EM

The idea is that we can regard $p(w|\theta_F^{(n)})$ as an estimated relevance distribution, $tf(w, \mathcal{F})$ as a kind of mixture distribution, and $p(w|C)$ as a kind of irrelevance distribution. Then, Eq. 33 fits Eq. 1, and Eq. 34 is the same distribution separation process as the Eq. 2, where $\hat{l}(R, I_{\overline{S}})$ is the estimated relevance distribution. It further demonstrates that the EM iterative steps in MMF can actually be simplified to the linear separation solution similar to Eq. 2.

Another simplified solution to MMF was proposed in.⁵ This solution is derived by Lagrange multiplier method, and the complexity of its divide & conquer algorithm is $O(n)$ (on average) to $O(n^2)$ (the worst case). On the other hand, our simplified solution in Eq. 34 was analytically derived from the convergence condition of the EM method in the MMF approach, and the complexity of the linear combination algorithm in Eq. 34 is further reduced to a fixed linear complexity, i.e., $O(n)$.

Besides of providing a simplified solution with linear complexity to the EM method in MMF, DSM shows an essential difference regarding the coefficient λ . In MMF, the proportion of relevance model in the assumed mixture model $tf(w, \mathcal{F})$ is controlled by λ , which is a free parameter and is empirically assigned to a fixed value before running the EM algorithm. On the other hand, in DSM, as aforementioned in Section , λ for each query is estimated via an analytical procedure based on its lower bound analysis (see Section), a minimum correlation analysis (see Section), and a maximal KL-divergence analysis described in Section .

Experiments

We have theoretically described the relation between Mixture Model Feedback (MMF) approach and our DSM method. Experiments in this section provide empirical comparisons of these two methods.

MAP (chg% over MMF)	WSJ8792	AP8889	ROBUST2004	WT10G
MMF	0.3388 ($\lambda = 0.2$)	0.3774 ($\lambda = 0.2$)	0.2552 ($\lambda = 0.1$)	0.1282 ($\lambda = 0.3$)
DSM (λ fixed)	0.3386 ($\lambda = 0.2$)	0.3767 ($\lambda = 0.2$)	0.2487 ($\lambda = 0.1$)	0.1267 ($\lambda = 0.3$)
DSM-	0.3474(+2.54%)*	0.3870(+2.54%)*	0.2889(+13.21%)**	0.1735 (+35.34%)**
DSM	0.3565 (+5.22%)**	0.3915 (+3.74%)*	0.2957 (+15.87%)**	0.1735 (+35.34%)**

Statistically significant improvement over MMF at level 0.05(*) and 0.01(**).

Table 3. Comparison of DSM and MMF approach with TF and C as input distributions

As aforementioned, the EM iteration algorithm of MMF can be simplified as a distribution separation procedure (see Equation 34) whose inputs are two distributions $tf(w, \mathcal{F})$ (TF for short) and $p(w|C)$, where TF is the distribution for which the probability of a term is its frequency in feedback documents, and C is the collection distribution of terms. It has been shown that Equation 34 is actually a special DSM when TF and C are DSM's input distributions and λ is assigned empirically

without principled estimation, denoted as DSM (λ fixed) (see Table 3). Now, we compare MMF and DSM (λ fixed) to test Proposition 5 empirically.

At first, we directly measure the KL-divergence between resultant distributions of MMF and DSM (λ fixed). We report the results of Query 151 - 160 on WSJ8792 with $\lambda = 0.8$ in Figure 2, and results of other queries/datasets show similar trends. It can be observed from this figure that the KL divergence between the results of two mentioned methods, i.e., MMF and DSM (λ fixed) tends to be very close to zero, which supports the proof of their equivalence, i.e., Proposition 5.

Next, we compare the retrieval performance of MMF and DSM (λ fixed). For MMF, we set λ to the value with the best performance, and this optimal value is also used in DSM with λ fixed. Experimental results are shown in Table 3. We can find that performances of these two methods are very close, which is consistent with the analysis in Section . This result again confirms that the EM methods in MMF can be simplified by Equation 34 with little decline of performance.

In addition, we also apply DSM- and DSM in the same setting (i.e., when TF and C are input to DSM-/DSM as the mixture distribution and seed irrelevance distribution, respectively) for comparison. It is demonstrated in Table 3 that the performances of both DSM- and DSM are significantly better than MMF. This is because although MMF and DSM(λ fixed) empirically tune λ for each collection, the value of this parameter is the same for each query given the test collection. On the contrary, DSM- and DSM adopt the principled estimation of λ *adaptively* for each query based on lower-bound analysis, minimum correlation analysis and maximum KL-divergence analysis. This set of experiments demonstrate that the estimation method for λ in DSM method is crucial and effective for background noise elimination.

Conclusion

In this paper, we further study the theoretically properties of Distribution Separation Method (DSM). Specifically, we generalized the minimum correlation analysis in DSM to maximum (original and symmetrized) KL-divergence analysis. We also proved that the solution to the well-known Mixture Model Feedback can be simplified using the linear combination technique in DSM, and this is also empirically verified using standard TREC datasets. Equipped with these analysis, DSM now has solid theoretical foundation which makes its possible to further extend DSM with principle. In addition, comparison with MMF helps us to find more scenarios for application of DSM.

References

1. Zhang, P., Hou, Y. & Song, D. Approximating true relevance distribution from a mixture model based on irrelevance data. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, 107–114 (2009).
2. Zhai, C. & Lafferty, J. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, 403–410 (2001).
3. Rodgers, J. L. & Nicewander, A. W. Thirteen ways to look at the correlation coefficient. *The American Statistician* **42**, 59–66 (1988).
4. Zhai, C. A note on the expectation-maximization (em) algorithm. *Course note of CS410* (2007).
5. Zhang, Y. & Xu, W. Fast exact maximum likelihood estimation for mixture of language model. *Inf. Process. Manage.* **44**, 1076–1085 (2008). URL <http://dx.doi.org/10.1016/j.ipm.2007.12.003>.